DOCUMENT RESUME

ED 421 490 TM 028 449

AUTHOR De Champlain, Andre F.; Margolis, Melissa J.; Ross, Linette P.; Macmillan, Mary K.; Klass, Daniel J.

TITLE Setting Test-Level Standards for a Performance Assessment of Physicians' Clinical Skills: A Process Investigation.

PUB DATE 1998-04-00

NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Clinical Experience; Higher Education; Interrater Reliability; Medical Education; Medical Students; *Performance Based Assessment; *Physicians; *Skills; Standards; *Student Evaluation

IDENTIFIERS *Standard Setting; *Standardized Patients

ABSTRACT

The purpose of the present investigation was to address several critical issues relating to setting a performance standard on a nationally administered standardized patient examination (SPX). The specific goals of the study were to: (1) compare pass/fail rates from this exercise to those of past studies undertaken with the same examination; (2) assess inter-rater classification consistency at both the case and test levels; and (3) determine whether a partially or fully compensatory model more accurately accounts for the process by which expert judges set test-level standards. Examinee data rated by the standard-setting judges in this study were taken from the 1996-97 administration of the National Board of Medical Examiners standardized patient examination at two testing sites. A random sample of 160 fourth-year medical students was chosen for the investigation, and 8 clinicians were recruited to provide the expert judgments. This information will be valuable in helping determine how to best set a fair and defensible standard for future forms of the SPX. (Contains 4 tables, 1 figure, and 22 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 421 490

Setting Test-Level Standards for a Performance Assessment of Physicians' Clinical Skills: A Process Investigation

André F. De Champlain, Melissa J. Margolis, Linette P. Ross,

Mary K. Macmillan, & Daniel J. Klass

National Board of Medical Examiners

André De Champlain

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the meeting of the American Educational Research Association

San Diego, CA

Monday, April 13, 1998

TM028449

Abstract

The purpose of the present investigation was to address several critical issues relating to setting a performance standard on a nationally administered standardized patient examination (SPX). Specifically, the goals of this study were: (1) to compare pass/fail rates from this exercise to those of past studies undertaken with the same examination, (2) to assess inter-rater classification consistency at both the case and test levels, and (3) to determine whether a partially or fully compensatory model more accurately accounts for the process by which expert judges set test-level standards. This information will be valuable with respect to helping us determine how to best set a fair and defensible standard for future forms of our SPX.

Setting Test-Level Standards for a Performance Assessment of Physicians' Clinical Skills: A Process Investigation

A considerable amount of literature has been dedicated to issues relating to performance assessments and their use in medical education (Colliver & Williams, 1993; Swanson, Norman, & Linn, 1995; van der Vleuten & Swanson, 1990). One particular type of performance assessment that has been used with increasing frequency to assess the clinical skills of medical students is the standardized patient (SP) examination (Vu & Barrows, 1994). SPs are laypersons trained to depict clinical encounters in a consistent, standardized fashion (Barrows, 1987).

Although there is no shortage of research assessing the reliability of SP examination scores (Swanson & Norcini, 1989; Swanson, Norman, & Linn, 1995; van der Vleuten, 1996; Vu & Barrows, 1994), relatively little research has focused on addressing other important measurement issues. One of these critical issues is how to best set a performance standard for these alternative assessments.

Most of the methods that have been proposed to set standards for performance-based examinations are extensions of procedures originally devised for use with multiple-choice items (Livingston & Zieky, 1982). Some authors have proposed modifications of the Angoff (Hambleton & Plake, 1995; Mittal, Chai, & Wennberg; Norcini *et al.*, 1993) and contrasting-group approaches (Clauser & Nungester, 1996; Colliver, Barnhart, Marcy, & Verhulst, 1994; Dauphinee, Blackmore, Smee, Rothman, & Reznick, 1997; Ross, Clauser, & Clyman, 1994). Others have advocated using polytomous item response theory models in the standard setting process for performance assessments (Luecht, 1993). Jaeger (1995) states that these procedures, as they were originally intended to be used, assume that the latent proficiency space is unidimensional and that the items comprising an examination contribute to a summative scale. These assumptions are likely to be violated with complex performance-based measures and could account for some of the mixed results reported in past studies. In response to this, Jaeger (1995) and Putnam, Pence, and Jaeger (1995) proposed judgmental policy capturing (JPC) and dominant profile (DP) approaches for standard setting. These methods seem better able to capture the complex, multidimensional nature of alternative assessments by allowing judges to review entire performance profiles. JPC and DP methods typically entail approximating standard setting policies of judges using regression modeling.

More importantly, JPC and DP methods enable the researcher to examine the process by which a pass-fail decision is reached and to determine which type of model most aptly captures the policies of judges. In other words, does a fully compensatory model adequately reflect the standard setting process (i.e., can an examinee make up for poor performance on one case by exemplary performance on another case) or are certain cases essential in order to obtain an overall passing grade (non-compensatory model)? Jaeger (1995) and Putnam, Pence, and Jaeger (1995) proposed a JPC-based approach to standard setting that involves multiple iterations and intermediate feedback to judges. This process enables panelists to articulate their policies as clearly as possible.

Clauser, Clyman, Margolis, and Ross (1996) and Ross, Clauser, Margolis, Orr, and Klass (1996) also completed studies that were aimed at better understanding the process by which expert judges set standards on patient management (PM) and clinical skills assessments (CSA). For the CSA, a fully compensatory model was adequate in mapping case-level judgments to test-level decisions. However, for the PM examination, a partially compensatory model seemed to account for examination-level policies. That is, certain cases were deemed critical in order to receive a passing score at the test level.

The purpose of the present investigation was to replicate and extend the research of Ross *et al.* (1996) with a nationally administered standardized patient examination. Specifically, the goals of this study were: (1) to compare pass/fail rates from this exercise to those of past studies, (2) to assess inter-rater classification consistency at both the case and test levels, and (3) to determine whether a partially or fully compensatory model more accurately accounts for the process by which expert judges set test-level standards.

Methods

The National Board of Medical Examiners' (NBME) SP examination is designed to assess the clinical skills of medical students who are about to enter their first postgraduate year. In this examination, examinees rotate through a series of clinical scenarios, or cases, and are evaluated on their history-taking (Hx), physical examination (PE), communication (CM) and interpersonal (IP) skills. The first three skills are assessed using case-specific checklists which are composed at most of 25 dichotomously-scored items that indicate whether or not a student has

completed a specific task. Two of the three skills are typically assessed per case. Interpersonal skills are assessed using the Patient Perception Questionnaire (PPQ), an inventory that is identical across cases and consists of six items scored on a five-point Likert scale. The checklist and PPQ are both completed by the SP after each examinee-patient encounter.

The examinee data rated by the standard-setting judges in this study were taken from the 1996-1997 administration of the NBME SP examination at two testing sites. A random sample of 160 fourth-year medical students who had taken the full 12-case SP examination was chosen for this investigation. A subset of six cases constituting a representative sample of all cases outlined in the test blueprint (with respect to skill and content specifications), was selected for the standard setting exercise.

Eight clinicians were recruited to provide the expert judgments (the clinicians will heretofore be referred to as "judges"). They practiced in a large number of medical specialties and all had teaching and/or supervisory responsibilities for fourth-year medical students. The group of eight judges was first oriented to the overall test with a brief overview of the test blueprint, training protocols and cases. The group then reviewed the first case by viewing a videotaped performance exemplar and discussing the specifics of the case.

After the case review, judges were divided into two groups of four and asked to independently rate the adequacy of 10 practice examinee performance profiles. These profiles contained scored checklist and PPQ items for each examinee, by case. The judges were told to first rate each skill within the case (e.g., Hx, PE and IP) and then to provide a case-level rating. The rating scale ranged from "1" (clearly inadequate performance) to "5" (clearly adequate performance), and it was emphasized that a rating of "3" (just adequate) constituted a passing score. The judges rated the practice performances and then discussed their ratings, thus allowing them to express their policies and to hear how other judges were attributing ratings. Judges were then asked to rate the performances of the remaining 150 examinees. After completing the ratings for each case the judges reviewed the information and videotape for the next case. This process was repeated for each of the six cases. After completing the case-level ratings for all cases, judges were asked to independently review their individual skill-within-case and case-level ratings and to then provide a global examination rating for each student (using the same 1-5 scale). The

ratings of the 150 examinees in the actual rating sessions (and not those of the 10 practice examinees) were used in all analyses.

For our analyses, a case-level "pass" was assigned if the majority of judges (five or more of the eight) gave a rating of "3" or more to an examinee's case performance. Similarly, at the test level, a "pass" was assigned if the majority of judges gave a rating of "3" or more to an overall test performance. Finally, average case ratings were calculated by simply computing the mean rating (across judges) for each case. The mean of the six averages was then computed and treated as the overall average case rating for a given examinee.

Results

Descriptive statistics summarizing pass/fail rates at both the case and test levels were computed and are presented in Table 1. Given that results were nearly identical across each group of four judges, findings will be collapsed and reported across all eight clinicians.

Insert Table 1 about here

At the case level, overall pass rates ranged from .56 (Case 6) to .97 (Case 5). Ninety-two percent of examinees received a passing score at the test level. These results are consistent with those reported in a previous study (Ross *et al.*,1996).

Inter-judge classification consistency was assessed by computing uncorrected as well as (chance) corrected proportions of agreement for each pair of judges. Results are provided in Table 2.

Insert Table 2 about here

Uncorrected proportions of agreement ranged from .73 (Case 6) to .95 (Case 5). At the examination level, the mean proportion of agreement, uncorrected for chance, was .89. It was also of interest to assess the inter-judge

classification consistency that was beyond what would be expected by chance agreement alone. Adjusted inter-judge coefficients were therefore computed for all pairs of judges. In all instances, chance was set at a fixed value of .50 which corresponds to $1/n$, where n is the number of possible decision outcomes (i.e., pass/fail). At the case level, mean corrected coefficient values ranged from .46 (Case 6) to .90 (Case 5). At the examination level, the mean agreement rate was .77. Using Landis and Koch's (1977) guidelines, these coefficients (both uncorrected and adjusted), were indicative of very good to excellent agreement.

Generalizability analysis was used to further investigate the degree of consistency with which judges attributed ratings, both at the case and overall examination levels. A person by rater nested within group ($P \times R{:}G$) analysis was undertaken for each case and for the overall test-level decision. The resulting dependability coefficients are presented in Table 3.

Insert Table 3 about here

Dependability coefficients (Φ) for the six cases ranged from .87 in Case 6 to .96 in Case 3, with an average Φ of .92. This suggests that the use of eight judges, or two groups of four clinicians each, was sufficient to produce fairly reliable case-level judgments. The dependability coefficient for the total test ratings using eight judges was .86. It was also of interest to see what specific impact increasing the number of judges would have on the resulting dependability coefficients. The latter values are shown in Table 3. The estimated dependability coefficients based on 12 judges ranged from .89 to .97, with an average Φ of .93 . The range for 16 judges was from .90 to .98 with an average Φ of .94. For 20 judges, the range was .90 to .98, with an average Φ of .95. Finally, for 24 judges, the range was again .90 to .98 with an average Φ of .95. At the test decision-level, dependability coefficient values respectively corresponded to .90, .91, .93, and .93 when the number of judges was increased to 12, 16, 20 and 24.

It was also of interest to examine how judges map case-level judgments to test-level decisions. Figure 1 provides a plot of the overall average case rating by the number of cases passed.

Insert Figure 1 about here

This figure is useful to help determine whether a partially compensatory or fully compensatory model accounts for the judges' policies when making test-level decisions. The average case rating is plotted along the *x*-axis whereas the *y*-axis shows the number of cases passed. Passers (squares) and failers (diamonds) are also differentiated on this plot. A vertical line separating passers from failers would suggest that a fully compensatory model is used by judges when setting a standard. That is, doing poorly on one case can be compensated for by doing well on other cases. Alternatively, a horizontal line separating passing from failing performances would suggest that a non-compensatory model is being used when aggregating case-level decisions to arrive at a test-level decision. In other words, a certain number of cases must be passed in order to pass the test overall. Looking at Figure 1, it appears as though the fit of a compensatory model is best. That is, a student can still receive an overall "pass" rating even if he or she did poorly on one case as long as the performance on other cases is strong.

Another way of determining which model most accurately reflects the process by which judges use case-level decisions when attributing examination-level ratings is to use logistic regression analysis. The independent variables in the model corresponded to the six case-level decisions (pass/fail or 1/0) and the overall test-level decision was treated as the dependent variable. The fit of the model was examined in addition to the estimate of the case odds ratios. These odds ratios will provide important information regarding the relative contribution of each case towards the test-level decision. Table 4 presents the results of the logistic regression analysis.

Insert Table 4 about here

The parameter estimates and their respective *p*-values indicate which cases were significant in predicting the overall pass-fail decision. Using a *p*-value of .10 or less as being indicative of a significant effect, all cases (with the exception of Case 3), were important in predicting the overall pass-fail decision. Looking at the odds ratios for

the various case-level decisions, it appears that the overall probability of passing at the test level is a function of passing three particular cases: Cases 2, 5 and 6. In fact, the odds of obtaining a "pass" rating at the test level are 14, 18, and 15 times greater if one passes Cases 2, 5, and 6, respectively. This suggests that a student who passes Cases 2, 5 and 6 can probably still pass the test even if he or she did poorly on other cases. This is further evidence for a compensatory model.

Discussion

Results obtained in the present investigation parallel those reported in a previous study focusing on standard setting for the NBME SP examination (e.g., Ross *et al.*, 1996). Overall pass rates at both the case and examination level were nearly identical to those described with another cohort of examinees. It should be pointed out that the relatively low pass rates obtained for cases two and six were not entirely unexpected. The latter seem to reflect an unfamiliarity (on the part of the student), with the physical examination procedures that were deemed necessary for an adequate performance on these two cases. These results are possibly attributable to their lack of training in the use of these maneuvers at the time of testing. It was also encouraging to note that the overall pass rate of .92 was similar to that reported for first-time USMLE STEP2 test takers, a population nearly identical to the one targeted by our SP examination. Additionally, both NBME SP and STEP2 examination scores have been shown to be moderately correlated (Clauser, Ripkey, Fletcher, King, Klass, & Orr, 1993).

Similar case and overall classification consistency rates were also noted among judges who participated in our standard setting exercise. There nonetheless appeared to be a substantial amount of variation with respect to corrected and uncorrected agreement pass/fail agreement rates for Cases II and VI, as evidenced by the larger standard deviations and ranges among ratings. This would seem to suggest the presence of an outlying or overly stringent ("hawk") judge. Additional training, including performance exemplars, would seem warranted for these two cases in future standard setting exercises.

Results from the *G*-study indicate that the panel of eight clinicians was able to reliably make case-level judgments. The costs associated with increasing the number of judges beyond 12 do not seem warranted given the

modest increases in dependability coefficient values. For example, doubling the number of judges from eight to 16 would result in an increase of .05 in the dependability coefficient value for an examination-level decision. However, the feasibility of using alternative measurement designs should be explored in ulterior Decision (*D*) studies. In particular, nested designs (which would obviate the need for all judges to rate all student performances), might prove to be cost effective while maintaining an adequate level of reliability. Similar research focusing on the post-encounter note included in this examination suggests that using groups of four raters per case yields dependability indices that are nearly identical to those obtained with eight judges in a fully-crossed design (De Champlain, Fletcher, Klass, Margolis, 1998).

Examination of the test-level policy adopted by our standard setting committee indicates that clinicians seem to be using a partially compensatory model for setting the pass/fail standard. The odds ratios estimated when attempting to predict the examination-level decision as a function of the six-case decisions indicate that an overall pass seems to be highly dependent upon doing well on three cases in particular. That is, strong student performances on the latter three cases might be able to compensate for weaker efforts on the remaining three cases. This finding is again comparable to that reported by Ross *et al.* (1996) in a previous investigation.

Although encouraging, the results presented in this study should be interpreted in light of several caveats. First, the standard setting exercise was based on a restricted sample of cases (about half the number that would be included in an actual test form) and should be replicated with different and longer case sets. This research is currently being undertaken with our examination. Secondly, it would be important to cross-validate our results with different standard setting committees. Although we attempted to do this by dividing our clinicians into two equal groups, the discrepancies noted in agreement rates on two cases would suggest that additional investigations are needed. Finally, it would be useful to present our findings to the eight participating judges in order to obtain their reactions, especially with respect to the section pertaining to the modeling of their standard setting policies. Hambleton and Plake (1995) showed that a discrepancy sometimes exists between the policy that judges believe they are employing when making pass/fail decisions and the one suggested by empirical analyses.

Nonetheless, the results obtained in this and other ongoing studies continue to be of great use to the present testing program. It is hoped that other researchers will benefit from these findings and that future efforts will contribute to establishing fair and defensible standards for this type of examination.

References

Barrows, H.S. (1987). *Simulated (Standardized) patients and other human simulations: A comprehensive guide to their training and use in teaching and evaluation.* Chapel Hill, NC: Health Sciences Consortium.

Clauser, B.E., Clyman, S.G., Margolis, M.J., & Ross, L.P. Are fully compensatory models appropriate for setting standards on performance assessments of clinical skills? (1996). *Academic Medicine, 71*, s90-s92.

Clauser, B.E., & Nungester, R.J. (1996). Setting standards on performance assessments of physicians' clinical skills using contrasting groups and receiver operating characteristic curves. *Evaluation & the Health Professions, 19*, 516-539.

Clauser, B.E., Ripkey, D., Fletcher, B., King, A.M., Klass, D.J., & Orr, N. (1993). A comparison of pass/fail classifications made with scores form the NBME standardized-patient examination and part I examination. *Academic Medicine, 68*, s7-s9.

Colliver, J.A., Barnhart, A.J., Marcy, M.L., & Verhulst, S.J. Using a receiver characteristic (ROC) analysis to set passing standards for a standardized-patient examination of clinical competence. (1994). *Academic Medicine, 69*, s37-s39.

Colliver, J.A., & Williams, R.G. (1993). Technical issues: Test application. *Academic Medicine, 68*, 454-460.

Dauphinee, W.D., Blackmore, D.E., Smee, S., Rothman, A.I., & Reznick, R. (1997). Using the judgments of physician examiners in setting the standards for a national multi-center high stakes OSCE. *Advances in the Health Sciences Education, 2*, 201-211.

De Champlain, A.F., Macmillan, M.K., Klass, D.J., & Margolis, M.J. (1998, July). *Assessing the reliability of post-encounter note scores in a large-scale standardized patient examination: Comparing the scoring consistency of medical chart abstractors and physicians.* Paper to be presented at the Eight International Ottawa Conference on Medical Education and Assessment, Philadelphia, PA.

Hambleton, R.K., & Plake, B.S. Using an extended Angoff procedure to set standards on complex performance assessments. (1995). *Applied Measurement in Education, 8*, 41-45.

Jaeger, R.M. Setting standards for complex performances: an iterative, judgmental policy-capturing strategy. (1995). *Educational Measurement: Issues and Practice, 14*, 16-20.

Landis, J.R., Koch, G.G. The measurement of observer agreement for categorical data. (1977). *Biometrics, 33*, 159-173.

Livingstone, S.A., & Zieky, M.J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Mittal, N., Chai, S., & Wennerberg, S. (1996, April). *Comparison of standard-setting methods for essay performance assessments*. Paper presented at the meeting of the American Educational Research Association, New York, NY.

Norcini, J.J., Stillman, P.L., Sutnick, A.I., Regan, M.B., Haley, H.A., & Williams, R.G. Scoring and standard setting with standardized patients. (1993). *Evaluation and the Health Professions, 16*, 322-332.

Norcini, J.J., & Swanson, D.B. Factors influencing testing time requirements for measurements using written simulations. (1989). *Teaching and Learning in Medicine, 1*, 85-91.

Putnam, S.E., Pence, P., & Jaeger, R.M. A multi-stage dominant profile method for setting standards on complex performance assessments. (1995). *Applied Measurement in Education, 8*, 57-83.

Ross, L.P., Clauser, B.E., & Clyman, S.G. (1994). A comparison of two methods for establishing case level standards for performance assessments. In *Proceedings from the Sixth Ottawa Conference on Medical Education* (pp.235-238). Toronto, Ont.: Faculty of Medical Education.

Ross, L.P., Clauser, B.E., Margolis, M.J., Orr N., & Klass D.J. An expert-judgment approach to setting standards for a standardized-patient examination. (1996). *Academic Medicine, 71*, s4-s6.

Swanson, D.B., Norman, G.R., & Linn, R.L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher, 24*,11-35.

van der Vleuten, C.P.M. The assessment of professional competence: Developments, research and practical implications. (1996). *Advances in Health Sciences Education, 1*, 41-67.

Van der Vleuten, C.P.M., & Swanson, D.B. Assessment of clinical skills with standardized patients: State of the art. (1990). *Teaching and Learning in Medicine, 2*, 58-76.

Vu, N.V., & Barrows, H.S. Use of standardized patients in clinical assessments: Recent developments and measurement findings. (1994). *Educational Researcher, 23*, 23-30.

Table 1

Case- Test-Level Pass Rate Descriptive Statistics

	Proportion of Passing Performances: Individual Judges		
Case	Overall	Minimum	Maximum
1	.86	.73	.95
2	.67	.57	.93
3	.89	.87	.93
4	.87	.81	.90
5	.97	.87	.99
6	.56	.29	.82
Total	.92	.83	.98

Table 2

Case-Level and Test-Level Classification Consistency Descriptive Statistics

	Proportion of Agreement Among Judges Uncorrected for Chance			Proportion of Agreement Among Judges Corrected for Chance (chance = 0.5)		
Case	Mean (SD)	Minimum	Maximum	Mean(SD)	Minimum	Maximum
1	.84(.05)	.75	.95	.67(.09)	.50	.90
2	.77(.09)	.65	.96	.53(.17)	.30	.92
3	.92(.02)	.87	.96	.85(.04)	.74	.92
4	.90(.04)	.82	.95	.81(.07)	.64	.90
5	.95(.04)	.88	.99	.90(.07)	.76	.98
6 *	.73(.12)	.44	.97	.46(.24)	-.12	.94
Total	.89(.03)	.85	.95	.77(.06)	.70	.90

Table 3

Estimated Dependability Coefficients (Φ) for Exercises Using Eight, 12, 16, 20 and 24 Judges and for the Total Test

	Number of Judges				
Case	8	12	16	20	24
1	**.91**	.93	.94	.95	.96
2	**.91**	.91	.92	.93	.93
3	**.96**	.97	.98	.98	.98
4	**.95**	.97	.97	.98	.98
5	**.90**	.93	.95	.96	.96
6	**.87**	.89	.90	.90	.90
Total Test	**.86**	.90	.91	.93	.93

Table 4

Logistic Regression Parameter and Odds Ratio Estimates

Predictor	Parameter estimate	*p*-value	Odds Ratio
Case 1 Decision	1.73	.075	5.62
Case 2 Decision	**2.64**	**.021**	**14.04**
Case 3 Decision	1.26	.300	3.52
Case 4 Decision	1.98	.037	7.21
Case 5 Decision	**2.87**	**.086**	**17.63**
Case 6 Decision	**2.74**	**.029**	**15.54**

Figure 1: Average Case Rating by Number of Cases Passed